1995

13. Future Psychometric Practices In Licensure Testing

Steven S. Nettles
Applied Measurement Professionals, Inc.

13

FUTURE PSYCHOMETRIC PRACTICES IN LICENSURE TESTING

Steven S. Nettles

Applied Measurement Professionals, Inc.

New technologies continue to emerge each year, and influence testing practices. In particular, in the last 10 years the personal computer has evolved from a curious and minimally useful tool to an indispensable partner in many certification and licensure testing programs. It is involved in every aspect—including candidate scheduling, test assembly, test administration, test scoring and analysis, and score reporting. Initially, it is used to determine the content to be included in the job analysis instrument, and later, to analyze the returned surveys. After the job analysis is completed and test specifications prepared, it can be used to bank test items written to the specifications. Assembly of test forms, and typesetting of final copy prior to printing can be expertly accomplished. When paired to an optical mark reader scanner, it can be used to score and analyze tests. As an alternative to paper-and-pencil test delivery, items can be loaded onto a computer and administered in a variety of alternate forms and can provide instantaneous feedback to candidates. Likewise, score reports can be prepared and mailed to candidates using information stored in the candidate database.

As the personal computer has gained in power, it has had significant impact on the psychometric practices of testing. Statistical packages written for the "PC" platform are now as powerful as their mainframe counterparts. This has increased the accessibility of resource hungry technologies such as Item Response Theory (IRT), making them available to many more individuals than those at universities and large testing companies. In turn, this availability has stimulated the research on new technologies, and encouraged their transition from "ivory tower" applica-

tions to real world, applied testing environments. Although the transition has not been totally painless, the initial trepidation has been overcome, and many organizations are beyond "testing the waters." They are in the operational mode of running IRT and classical psychometric test analyses concurrently. In this chapter, I will discuss what I consider to be the most significant of these technologies, as they relate to the major areas of testing, and attempt to forecast their impact on several areas of licensure testing practices throughout the 1990s.

JOB ANALYSIS AND TEST SPECIFICATIONS

Job analysis is the initial step in any well-designed licensure testing program. The purpose of job analysis is to identify the content to be included on the examination, commonly referred to as test specifications, thereby establishing content validity. A typical procedure includes the development of a sufficient number of task and/or knowledge/skill/ability (KSA) statements that totally describe the important job activities. These statement are then subjected to evaluation by a group of job experts in which the most important activities are identified through a rating process. The rating results are used to develop test specifications—the content areas to be covered on the examination and their relative emphasis. A common procedure involves a committee of job experts making rational decisions about the structure and relative weighting of the content. For example, the structure may be defined as three major content areas, and the relative weighting may be 20% for Content Area I, 35% for Area II, and 45% for Area III.

Several methods exist for making these determinations statistically. However, not all have a sound empirical basis. Specifications are sometimes determined by initially combining several rating scales together for each activity statement to determine a "criticality value." For example, in a job analysis study of law enforcement special agents, Sistrunk and Smith (1982) calculated a "Task Importance Value" by multiplying the difficulty and criticality ratings together and then adding the time spent rating to this product. Test section weights are sometimes calculated by summing individual criticality values for all task/KSA statements determined to be in that section. Although this procedure may have intuitive appeal, it has no more statistical basis than the rational approach described earlier. Although both rational and empirical procedures may yield the same results, it has been my experience that a carefully conducted rational judgement procedure produces very usable test specifications.

Rosenfeld and Thornton (1978) were among the first to use a more sophisticated statistical approach in job analysis in an occupational testing setting. To develop an interim task list, existing job descriptions were reviewed, and interview and observation techniques were used. The resulting task list was reviewed and revised in several states by advisory committees. This version was pilot tested prior to preparation of the final instrument. The task list was mailed to a large number of incumbents in all participating states for evaluation using several rating scales. Principal component factor analysis was used to verify the rational groupings of tasks into a smaller number of dimensions. Similarly, hierarchical cluster analysis was used to group incumbents who reported similar patterns of time usage. The

results indicated that the factor analysis groupings confirmed the rational groupings. The authors attributed this to the extensive review and revision that was undertaken in the development phase. The cluster analysis revealed nine clusters of incumbents, some of whom were performing more specialized duties. The major job dimensions were linked to cognitive abilities by both measurement experts in a group session, and job incumbents and their supervisors through the mail with extensive directions. The authors concluded that the most preferable way to accomplish this linking was in a group session with measurement experts directing job experts in the process.

Shaefer, Raymond, and White (1993) evaluated the efficacy of two different statistical strategies, cluster analysis and multidimensional scaling (MDS), and two different rating scales, frequency and similarity, for establishing test specifications. Task frequency ratings and task similarity ratings were collected on a sample of 125 tasks for emergency nurses. Cluster analysis was used as the primary procedure, with MDS used for interpretation for both scales. The authors determined that the results based on similarity ratings, as opposed to frequency ratings, were more useful and interpretable. However, they do not recommend discarding frequency ratings, as they may be useful in helping to organize traditional multiple-choice examinations, and provide insight into another dimension of content description. A further caution is offered in that the results are based on the study of an occupation that may be more homogenous in terms of work activities than other occupations. Despite these caveats, this study provides a promising direction for future studies to pursue when empirical data are desired to supplement domain specifications based on expert committee judgement.

A common procedure in establishing test specifications is the use of a taxonomy or typology for item classification within content area as an additional level of specificity. The rationale is that because differing cognitive demands are required for the successful performance of the required job activities, test specifications should reflect the cognitive demands of the target job. For example, medical laboratory technologists are required to collect tissue samples and evaluate them for various abnormal conditions. Because collecting requires a different cognitive level than evaluating, items written to assess the former should be written at a different cognitive level than the latter. In Bloom's taxonomy (Bloom, Englehart, Furst, Hill, & Krathwohl, 1956) nomenclature, "collecting" items would be written at the application level and "evaluating" items would be written at the analysis/evaluation level. This classification appears intuitive. However, after assisting numerous expert examination committees in the performance of item review and revision, obtaining unanimous agreement among them on the particular classification of a particular item is often difficult. Although some believe that such an acceptable classification system does not exist (see Haladyna, 1992a), test specifications using a cognitive level system can result in an examination with additional evidence in support of content validity.

Job analysis is an area in which existing statistical techniques will represent the "new technologies" that will be applied to job analysis data. Expert judgement will continue to be used, but will be supplemented with empirical techniques such as multivariate analyses. As a result, the commonly reported descriptive data may

have additional empirical evidence to support expert committee judgement. As the above studies indicate, the application of multivariate techniques to supplement the interpretation of descriptive statistics at the unit level promises a new direction in job analysis research.

ITEM FORMATS

After job analysis has been completed, a multiple-choice examination is frequently developed to assess the important content domains. The development of high quality test items and their format is the next step. Research in item format has been cyclical, but lately is an area that has drawn increased attention. Downing (1992) investigated true-false and alternate-choice multiple-choice question (MCQ) formats. The alternate-choice format is essentially a two option MCQ. When compared to the traditional simple MCQ, the advantages of these formats include greater ease of writing, and the presentation of more items to the examinee in a similar period of time. The disadvantages are that both formats are likely to result in inaccurate candidate scores because of guessing, and that true-false items may be subject to ambiguity, as many items may not be completely true or false. Downing concludes that the alternate-choice format may be appropriate in some situations for credentialing (and by extension, licensure) examinations.

Haladyna (1992b) studied various multiple choice question formats, including alternate-choice (AC), true-false (TF), complex multiple-choice (CMC) of which K-type is a subset, multiple true-false (MTF), and context-dependent item set (CDIS). In the CMC format, several potentially correct statements are presented, followed by various combinations of those statements. The MTF format is similar to the CMC, except that candidates are allowed to respond to each of the statements with either a true or false. He concludes that the CMC format should be discontinued, and that the MTF be used in its place. He feels that both the MTF and the CDIS format can be used to objectively score complex cognitive behavior efficiently.

In Haladyna (1992a), context-dependent item formats were examined exclusively. One caution on context-dependent items is that the items should be independent, so that the candidate is not penalized more than once for a wrong answer. An exception to this is in patient management problems (PMP). In PMPs candidates are presented a series of scenarios in which they are asked to gather information, process it, and select a course of action (Hixon, 1985). Provisions are made for those candidates who select an inappropriate course of action, by redirecting them to the proper path.

That CMCs not be used is congruent with Albanese (1993) in which several studies on CMCs in general, and Type K items in particular, were reviewed. Type K items present four primary statements, whereas the options are a fixed set of five combinations of the primary statements (Hubbard, 1978). Type K items were observed to have more clueing that leads to increased scores, decreased reliability, and are more likely to be deleted at key verification. However, he concluded that few studies have been done on the more general format of the CMC, and it may address some of the problems presented for the Type K format.

In other studies, it was concluded that although reliability was similar for the CMC and simple multiple-choice formats, candidates respond to fewer CMC format items in the same time period (Dryden & Frisbie, 1975), and that candidate scores on a CMC test represented a mixture of knowledge, test wiseness, and blind guessing (Kolstad, Bryant, & Kolstad, 1983). Studies by Case and Downing (1989), and Dawson-Saunders, Nungester, and Downing (1989) provide additional evidence in support of their discontinuance.

However, the results of these studies are contradicted by Nettles (1987), in which the psychometric characteristics of simple multiple-choice (SMC) and CMC items were compared. Data were collected from 3,500 individuals who had taken a self-assessment examination for a large allied health profession. In comparison to simple multiple-choice items, CMC items were found to fit the three-parameter IRT model equally well. Also, in evaluating the amount of information in the wrong options, both were identified in proportions comparable to their actual representation on the test. Additional unpublished studies using IRT three-parameter (3-PL) methodology conducted on a certification test for one allied health profession indicate that both item types are comparable in discriminating power and amount of information, as well as difficulty and guessing indices (see Table 1). Support was found for the other studies' observations that, in general, CMC items tend to be more difficult than SMC items. The one exception is that SMC items involving calculations (math items) were observed to have the lowest mean *p*-value.

However, another unpublished study conducted for a different allied health licensing test presented conflicting results. This study indicated support for the earlier conclusion by others that CMC items tend to be more difficult (again, excluding math items) and do not discriminate as well as SMC items. The other interesting finding was that negatively worded items were equal to positive items in discrimination and difficulty (see Table 2). This result is in conflict with other studies (see below), which have recommended against the use of negatively worded items due to their poor psychometric properties.

Table 1. Mean Item Statistics by Item Type for Group A.

Type	*P*-value	Point-biserial	a	b	c
SMC positive (n=103)	.75	.26	.46	-2.0	.14
SMC negative (n=6)	.73	.19	.31	-1.6	.15
SMC calculation (n=3)	.53	.26	.43	.3	.10
SMC data table (n=10)	.67	.25	.46	-1.0	.13
CMC positive (n=18)	.71	.29	.47	-1.2	.14

Table 2. Mean Item Statistics by Item Type for Group B.

Type	P-value	Point-biserial	a	b	c
SMC positive (n=152)	.80	.25	.52	-1.6	.24
SMC negative (n=60)	.77	.28	.52	-1.4	.23
SMC calculation (n=20)	.74	.30	.58	-1.0	.21
SMC situational set (n=22)	.75	.28	.58	-1.0	.24
CMC data table (n=3)	.67	.26	.49	-0.5	.23
CMC positive (n=39)	.71	.22	.46	-0.8	.25

The data tend to support the recommendation against the use of the specialized CMC format, the K-type item. However, the jury is still out regarding the more general format. Perhaps additional studies will show the more general CMC format to be a valuable item type. One area in which the general CMC format has great utility is in the rewriting of negatively worded items, eliminating the "except" or "not."

In general, evidence does exist for strong support in recommending against the use of negatively worded items. Negatively worded items include words in the stem such as "except," "not," "least," or "false." Harasym, Price, Brandt, Violato, and Lorscheider (1992) found that although negatively worded items are easier to write, candidates tend to find them more difficult to read and interpret correctly. These findings are somewhat supported in unpublished studies conducted by Nettles on tests constructed for purposes of licensing and certification. As Table 1 indicates, negatively worded items were found to be the least discriminatory in one study, but equal to positively worded items in another study (see Table 2) in which negatively worded items appear to be equal to positively worded items in average discrimination using both classical and IRT statistics. Anecdotally, in numerous item review meetings conducted with expert committees, some committee members invariably miss the "not" or "except" when reading this type of item, and provide inappropriate suggestions for revision. My prediction for negative items is that additional studies will support the recommendation against their use.

Research continues on the optimal number of options. Lord (1980, p. 112) indicated that three-option multiple-choice items were more appropriate for high ability candidates, whereas five-option items more suitable for lower ability candidates. Others have concluded that three-option items are easier to prepare, and more concepts can be tested due to decreased response time per question (Costin, 1970; Owen & Froman, 1987, cited in Landrum, Cashin, & Theis, 1993). Landrum, et al. (1993) composed alternate forms of an examination for an undergraduate psychology course, one with three-options and one with four-options. They found

that the students scored higher on three-option as opposed to four-option items. In addition, evidence was found that three-option tests may be more difficult, after correcting for guessing. Despite these somewhat encouraging results in support of the three-option multiple-choice item, until data are collected from certification and licensing examinee populations, a migration from four-option multiple-choice items will not occur quickly.

Currently, much interest has been directed toward "authentic assessment," commonly termed performance testing, or, more generically, assessment using constructed-response items. Wainer and Thissen (1993) characterize constructed-response items as "more difficult to score reliably and objectively, but [providing] a task that may have more systematic validity" (p.103). Oral examinations can be considered a form of constructed-response assessment, and have been frequently used in licensure and certification examinations. They often present substantial potential problems to the examining body, in the form of candidate scheduling, examiner equivalency, fatigue, and bias. However, they remain a popular format, especially in medical assessment. For example, Schweibert, Davis, and Jacocks (1992) evaluated data from oral examinations given for physician certification in board specialties. They found positive correlations with medical school grade-point average (GPA) and oral examinations for several medical specialties. Oral examinations will continue to be used, but because of their inherent problems with standardization from one examinee to another and high administrative costs, with decreasing frequency.

Additional studies will be done to examine alternative ways to score constructed-response tests. Bridgeman (1992) compared quantitative GRE items using a multiple-choice, paper-and-pencil open-ended format, and a computer-based open-end format. A specially designed answer sheet was used for the open-ended paper-and-pencil format, such that candidates could grid in their answers on a machine-readable sheet. Candidates used the keyboard to enter their answers for the computer version of the open-ended questions. Although differences were observed at the item level among the alternative formats, total test scores were found to be comparable. Further, all formats rank ordered the candidates similarly, and gender and ethnic differences were trivial or nonexistent. Correlational studies with other college grades and other tests revealed significant but not meaningful differences among the formats. Bridgeman concluded that although both the open-ended and multiple-choice formats will probably produce the same results, the open-ended format is more representative of the problems the candidates will face in real life situations. He suggests that both psychometric and nonpsychometric considerations be equally weighed in the decision to use the open-ended format in testing.

Another consideration in authentic assessment is the issue of which behaviors to include in the assessment exercise. In a typical performance assessment, from all important behaviors identified by the job analysis, only a few can be selected for inclusion because of time constraints. Thus, the assessment instrument samples only a small proportion of all possible behaviors. Shavelson, Baxter, and Gao (1993) used generalizability theory to examine this issue. They describe a performance assessment as consisting of a particular combination of all possible tasks,

occasions, raters, and measurement methods. Data taken from studies on California elementary students in math and science were analyzed using generalizability theory. The results from one part of their study indicated a large source of measurement error was due to the person x task interaction, indicating that the particular task sampled played a major role in students' performance scores. They concluded that this finding was consistent with other studies in that to obtain a measure of achievement that is generalizable, a large number of tasks is necessary. Based on their results, they speculated that, assuming 15 minutes per task, a total of 2.5 hours testing time would be necessary to obtain a generalizable measure of student achievement. Generalizabilty theory appears to be well suited for this type of research.

Authentic assessment measures are frequently combined with multiple-choice tests. Wainer and Thissen (1993) examined the most efficient way to combine scores from two different formats of measurement instruments. They examined possible scenarios of combining mixed-format tests using two graphic procedures. One procedure, the "ReliaMin," allows one to determine the amount of testing time needed to achieve equal reliabilities for each format. In their example, in order for a constructed response test to achieve the same reliability as a 75-minute multiple-choice chemistry test, 3 hours of testing time would be needed. More time would be necessary for an examination in a "softer science" such as arts and humanities.

They also developed a similar procedure, termed "ReliaBuck," that examines the resource expenditure (scoring costs) for equally reliable but different test formats. Again comparing a multiple-choice to a constructed-response format for a chemistry examination, they estimated that the costs for the constructed-response portion was 3,000 times more expensive than the multiple-choice format of the examination. As above, the costs associated with an arts or humanities test would be approximately three times more expensive again. They conclude that it does not appear to be economically practical to equalize the reliabilities of different components of mixed-format tests.

Perhaps the most desirable authentic assessment will be used in computer-based testing (CBT). CBT has already been applied to patient management problems (PMPs), and has demonstrated several desirable characteristics in comparison with the standard paper-and-pencil (PAP) format using latent image technology. Latent image test booklets use a special developer ink to expose the desired response text associated with the stimulus scenario. In latent image test booklets, the response text remains invisible until a special developer pen is applied. Thus, the candidate can be considered to be "constructing a response" by exposing the selected answer. The major drawback to the PAP approach is candidate advancement through the problem in an alternative manner to the specified path. Other problems include the lack of opportunity for the candidate to change his or her mind after exposing a response, and the appearance of "random" marks in the latent image area. This forces the scorer to determine if the candidate was attempting to gain an unfair advantage by discretely exposing a portion of the latent image, or if the mark was truly an accidental occurrence. Using CBT, the first problem is eliminated, in that the candidate progresses through the problem as

presented by the computer. Although CBT will not allow the candidate to change his or her mind about selecting a response, the candidate will have little support in indicating a response was exposed by accident, especially if the candidate is prompted to affirm his or her choices.

As computer technology advances, and as prices drop, CD-ROMs can be used to provide still or motion pictures to supplement the scenario text. However, the storage of many images as compared to a single image can be costly in terms of storage resources. It is encouraging that a study by Shea, Norcini, Baranowski, Langdon, and Popp (1992) found both formats sufficiently similar to justify the use of still pictures for credentialing examinations. In this study, the psychometric characteristics of still pictures versus motion pictures were examined. The results indicated that still pictures were both more reliable and more difficult than motion pictures, but that both formats were highly correlated with themselves and other types of performance measures.

In summary, research will continue to identify the "perfect" item types and modes of presentation. The multiple-choice item will continue to play a major role in licensure and certification testing, and possibly, with fewer than the four- and five-option format that is popular at present. Similarly, authentic assessment will play an ever increasing role in occupational assessment. However, it is apparent that inclusion of constructed-response items can be costly both psychometrically and practically. Perhaps one way to integrate this format into existing test programs in a practical way is to combine both formats using CBT. For example, the written stem of the item could be replaced with a video application, and the candidate could respond to video options presented in the multiple-choice format. Regardless, new and better ways will be found to use authentic assessment techniques that will overcome some of the psychometric and practical shortcomings presently observed, and make the behaviors required to answer test items more similar to the behaviors required to make decisions in real life.

STANDARD SETTING

Once a test is developed, and preferably before it is administered for the first time, a passing point must be determined. Although initially many licensing tests relied on norm referencing, the current generally accepted procedure is one in which the passing point is determined through an absolute standard procedure such as those described in Livingston and Zieky (1982), specifically, the Angoff (1971), Ebel (1972), and Nedelsky (1954) techniques.

Livingston and Zieky (1982) identified the following five steps that most absolute standard methods have in common:

1. Selecting the judges to render the ratings.
2. Defining the borderline or minimally competent practitioner.
3. Training the judges to use the selected procedure.
4. Collecting the judgments.
5. Summarizing the individual judgments to arrive at a passing score.

Selection of the judges is a crucial part of the standard setting process. In general they should be experienced job experts, representative of the candidate

population, so that a diversity of opinion and knowledge are represented. Jaeger (1991) identified several characteristics of an expert, including that they excel in their areas of expertise, they are able to perform domain-relevant tasks rapidly and correctly, they seem to be more aware of errors they might make, and that they are more accurate than novices in ascertaining the difficulty of a problem.

Knowing what characteristics constitute expertise, the next task for the measurement expert is to assemble a group of these individuals for a passing point study. The question is always asked, "How many judges are needed for the study?" The answer can be partially determined by evaluating the amount of error that is tolerable in the selected standard. Jaeger (1991) suggests that the number of judges can be determined by estimating a reasonable standard deviation (RSD) of recommended standards and the desired standard error of the mean (DSE), substituting these values in the equation for the standard error of the mean, and solving for n, where $n = (\text{RSD}/\text{DSE})^2$. In his example, 4.65 was selected for the RSD, and 1.3 for the DSE, resulting in a recommendation of 13 judges. It is encouraging that this value falls within the range of general rule of thumb of 10 to 20 judges.

Training of the judges is another crucial part of the standard setting process. This training includes direction in establishing the definition of the minimally competent practitioner (MCP), as well as the actual rating process. In defining minimal competence, Mills, Melican, and Ahluwalia (1991) suggest using the test specifications as a basis for identifying entry level skills and minimally acceptable levels for the entry-level practitioner. Concerning the actual rating process, Reid (1991) suggests beginning with a practice set of items that have item statistics available. Discussion is encouraged among raters, especially for those items with diverse ratings, with the hope that judges will reconsider their initial ratings in light of the group discussion. Additional training should be provided for specific item formats that tend to be more difficult for candidates, for example, negatively worded items and those involving calculations. Reid concludes his discussion by suggesting three criteria for evaluating the training of judges, namely that standard setting ratings should (a) be stable over time, (b) be consistent with relative difficulties of the items, and (c) reflect realistic expectations.

Many studies have been done comparing the various techniques (e.g., Andrew & Hecht, 1976; Poggio, Glasnapp, & Eros, 1981; Skakun & Kling, 1980). In most of these studies, differing results were obtained for the various methods, although different groups of judges were used for each method. In general, the Ebel and Angoff procedures tend to establish higher passing points than the Nedelsky. However, Mills (1983) found agreement among three different methods. He compared the Angoff, the contrasting groups method, and the borderline group method. He attributed the congruence of results to the fact that the same group of judges were used for all three methods.

Over the past few years, the original Angoff procedure, or a modification thereof, appears to be the most commonly used of the three. The reliability of this procedure was studied by Norcini and Shea (1992). They examined the reproducibility of a set of standards in two different scenarios. In one study, they found that

standards set by independent groups of experts using the same methodology (Angoff) and test content were similar. In another study, they found that similar standards were set by a subset of experts for the same test materials over 2 years elapsed time. These results are reassuring in that they indicate that the Angoff procedure appears to be quite reliable.

Once the data forming a passing score are collected, the results from each judge must be combined to produce a useful result. The most common procedure for establishing a passing score is to sum the average of the individual ratings across all items on the examination—equally weighting each item. Plake and Kane (1991) investigated two alternative approaches to combining the ratings by examining different types of error in setting a passing score. One alternative established the passing score based on the sampling variance of the average ratings. The other alternative established a passing score by selecting the best match between the judges' ratings and the actual proportion of minimally competent practitioners answering each item correctly. Using simulated data, they also varied the number of judges involved in the study (5 vs. 10) and the number of items in the examination (25 vs. 50). They observed that all three methods provided similar levels of accuracy, and that using more raters resulted in more precision. Slightly higher accuracy was found based for the 50-item test. They concluded that the traditional and simpler method of using the sum of the average judges' ratings should be the method of choice. This result is encouraging in that most Angoff studies arrive at a passing score in this manner. Also, the results indicate that the use of as many judges as practically possible is supported, and that the occasional necessity of discarding an item from the test form from which the study was conducted will probably have little practical significance on the resulting passing point.

Occasionally, the entire results of a standard setting procedure are unacceptable, because they result in a passing score that is either too high or too low. Breyer (1993) investigated this problem using the results of three hypothetical studies in which the Beuk (1984) adjustment was made. In the Beuk procedure, a compromise between an absolute method (Angoff), and a relative (norm-referenced) procedure is allowed. For example, the judges participate in an Angoff procedure, and are then asked to estimate pass rate of a group of first-time candidates for that examination. Breyer's results indicated that the Beuk procedure adjusts the cut score in favor of the judgments that have the most agreement (i.e., those judgments with the lowest standard deviation). It appears that the Beuk procedure may be useful in some situations occasionally encountered by the licensing test measurement professional. However, on a cautionary note, Geisinger (1991) suggests that the modification "procedures proposed Beuk and Hofstee [(1983)] are valiant first steps" (p. 21), but need to be better developed before they are fully endorsed.

The determination of a passing point remains a crucial part of the licensing examination process. I suspect the Angoff procedure will remain the most popular technique, and at least one study indicates support for employing the traditional procedure of summing the judges' ratings across items to determine the passing score. It is hoped future studies will occur that will provide additional empirical

support for the standards set by the Angoff and other absolute standards techniques, as well as provide additional information on existing procedures for modification of the results.

TEST AND ITEM ANALYSIS

Wainer (1990) provides both an enlightening and humorous history of "mental testing," tracing testing from several hundred years B.C., where a performance test was used to determine national affiliation, and in China where proficiency tests sampling a candidate's performance were used for candidates for political office. This testing system was continually refined until, in the 19th century, the British used it as their model for establishing the Indian civil service. The British system was used as the foundation for the U.S. Civil Service System in the late 1800s. The early days of psychometrics around the turn of the century allowed the transition from individualized to mass test administration. Military testing programs were the first to use mental tests on a large scale, mainly to support the war efforts of World Wars I and II. College admissions tests began in 1901 and closely paralleled the military testing programs though the 1950s. Both of these groups are responsible for the popularity of classical test theory that is so widely used by testing groups in the fields of licensing and certification. Classical test theory continues to provide much useful information for the vast majority of tests in use today.

Although classical test theory is a very powerful model on which to base test development and analysis, some of its shortcomings are significant. According to Hambleton and Swaminathan (1985), one of the major problems is that all statistics are relative to the group of examinees who took the test. That is, the item statistics will vary from test administration to test administration, especially if subsequent test administrations are conducted on groups of dissimilar examinees. Additionally, the discrimination index is affected by the spread in variability of examinees and the *p*-value of the item. Further, reliability is dependent on the standard deviation of the test, the *p*-values, and the item discriminations. Thus, item statistics are meaningful only if they are derived from highly similar tests given to highly similar populations of examinees.

Another shortcoming is that classical test theory provides no basis for determining how an examinee might perform when confronted with a test item. For example, we may know that a particular candidate is very able, and that a particular test item is moderately difficult. We can "guesstimate" that this particular candidate will probably answer the item correctly. However, if Item Response Theory (IRT) has been used, it is possible to make a precise estimate (in terms of probability) of how a particular candidate will perform to a particular item.

Finally, classical item statistics do not inform test developers about the location of maximum discriminating power of items on the total score continuum. This precludes constructing the test to examine very efficiently for a given range (e.g., around the cut score).

A comparison between IRT and Classical Test Theory (CTT) can be made. IRT statistics are provided and their nearest counterpart in classical test theory is provided below in Table 3.

Table 3. Classical Test and Item Response Theory Comparisons.

Classical Test Theory	Item Response Theory
p -value: can range from .00 to 1.00 (high *p*-values indicate easy items)	"b" parameter: typically range from -3.0 to +3.0 (high b values indicate hard items)
item discrimination: (e.g., point biserial correlation) typically range from -.30 to +.50	"a" parameter: typically range from 0 to 2.0 (high values indicate better discrimination)
nothing similar in classical, although 1/number of options is sometimes used as an estimate of the probability of guessing the right answer	"c" parameter, also known as the guessing parameter: typically varies from 0 to .25
total test score: a measure of achievement on the particular group of items on the test	theta (Θ): the scale used to describe an examinee's ability in IRT
reliability of test: an indication of the similarity of the content domain of the test. Although no definite standard exists, a target of .90 can be considered desirable.	test information curve (TIC): sum of individual item characteristic curves (ICCs). Items can be selected to provide maximum information at various points of the TIC (e.g., around the cut score)

The work of Birnbaum (1968), Lord and Novick (1968), Rasch (1960), and Wright (1968) stimulated the measurement community during the 1970s and 1980s to provide the necessary research that enabled Item Response Theory (IRT) to become as popular as it is today.

Item Response Theory (IRT) is a more powerful (and more complicated) model of test theory. It is also known as latent trait theory—test performance can be predicted in terms of underlying traits. For example, if an underlying trait for a clerical examination is good written communication, one of the knowledges assessed in the test may be punctuation. An IRT model specifies a relationship between the observable examinee test performance and the unobservable traits or abilities assumed to underlie test performance. A successful model provides a means of estimating scores for examinees on the underlying traits. The traits must be estimated from observable examinee performance on a set of items. This is known as calibrating the item pool.

IRT proposes that a single trait underlies examinee ability, and that the probability of an examinee's performance on a test item can be determined if the difficulty of the item and ability of the candidate is known. If the assumptions of IRT can be met for a particular set of items, the performance of two examinees can be compared even if they do not take the same set of items, and item statistics are comparable even if different groups of examinees are used in their calculation. These two properties are termed item-free ability estimates and sample-free parameter estimates (Hambleton, 1989). To have invariant item parameters is very desirable when building tests using a database of test items.

IRT has an item level orientation. IRT makes a definite statement about the probability of answering an item correctly and a test taker's ability. This relationship must be estimated through item calibration—item analysis is used to determine the item statistical parameter estimate. The major result of using IRT is that both

candidates and items are placed on the same scale of measurement. This feature allows use of the test to make definite predictions about examinee performance regardless of the test items presented to different examinees.

IRT provides a graphical interpretation of how well an item performs—the item characteristic curve (ICC) indicates the probability of an examinee's response based on his or her ability. The ICC is a plot of performance of an item against some measure of ability. This is usually a smooth nonlinear curve that is fitted to the data. Each item's ICC can be added to determine the Test Information Curve (TIC), a concept similar to reliability in classical test theory.

According to Hambleton and Swaminathan (1985), the characteristics of a properly fitting IRT model consist of the following:

1. Examinee performance on a test can be predicted in terms of one or more characteristics referred to as traits.
2. An IRT model specifies a relationship between observable examinee item performance and the traits or abilities assumed to underlie performance on the test.
3. Examinee scores on the underlying traits can be estimated.
4. The traits must be estimated from observable examinee performance on a set of test items.

Thus, a test properly calibrated using IRT has several useful features. Number one is that the item parameter estimates are independent of the group of examinees used from the population of examinees for whom the test was designed. Further, examinee ability estimates are independent of the particular choice of test items used from the population of items which were calibrated. That is, a different group of items (e.g., an alternate test form) can be used for different examinees, but their scores are directly comparable. Further, a model is provided that allows the matching of test items and candidate ability. Also, the precision of ability estimates are known for each examinee. Finally, test models do not require strictly parallel tests to determine reliability (Hambleton, 1989).

Because of these features, the characteristics of a test assembled using an item pool calibrated with IRT statistics are known before the test is given—the test information curve (TIC) can be used to determine the effect of each item and its impact on the total test. Additionally, the use of IRT allows pre-equating—the passing score of the test can be empirically determined prior to the administration of the test. This can be useful in situations where immediate feedback on candidate performance is desirable, for example, in computer-based test administration.

One of the areas in which IRT can play a significant part is in test construction, particularly item selection. Because the amount of information is available for each item at a specified difficulty level in a calibrated pool, items can be selected that best contribute to the total information described for the test. In three-parameter terminology, these items are typically ones that possess high discrimination (a) values and low guessing (c) values at the appropriate difficulty (b) value for the test. According to Lord (1980), the following steps are involved in test construction using IRT methodology. First, the desired test information curve is determined.

Then, items are selected to fill the area under the target information curve, filling the hard to fill areas first. As items are selected, the test information curve is calculated, with new items selected until the calculated test information curve closely approximates the target information curve. For licensing tests, the target information curve should be highly peaked near the passing score.

IRT should not be considered as a total replacement for classical test theory. Even when IRT has been determined appropriate for use, classical item statistics should continue to be used in conjunction with IRT. Classical statistics provide useful, easily understood information regarding test items, particularly information about the performance of each of the options. However, the additional use of IRT in examination development and scoring allows for significantly increased information being available regarding items and candidates in particular, and the test in general. Thus, the overall precision of measurement of the candidate population is increased, a most desirable characteristic of any testing program.

Practically speaking, it is important to remember that classical test theory is more easily understood by the testing consumer than is IRT. The typical examination committee is composed of job experts with little knowledge of testing. With a moderate amount of training, they can understand *p*-values and item discrimination indices, and their derivation. IRT statistics are not as intuitive, and it is considerably more difficult to explain their origin to lay persons. Popham (1993) recommends that we not expect the testing consumer to unthinkingly accept information from the IRT specialists. Part of our job as measurement experts is to present the necessary information about IRT in a comprehensible manner to the uninitiated. After having attempted to explain IRT to several examination committees, I can truly say that is easier said than done. Discussing comparisons between *p*-values and *b*s, item discrimination and *a*s, and guessing and *c*s is relatively straightforward. Explaining the math behind these item statistics is considerably more difficult. Nevertheless, IRT is an important technology that will continue to play an increasing role in licensure testing.

Although IRT does allow for multidimensional, linear, and polychotomous models, most licensing and certification programs at present use the undimensional, nonlinear, dichotomously scored response models. For example, both the National Council of State Boards of Nursing (NCSBN) and the Board of Registry (BOR) used one-parameter logistic (1-PL) IRT to calibrate their item pools as a necessary prerequisite to offering their examinations using computer-adaptive testing (CAT) technology. The NCSBN have implemented their CAT program, after several years of beta testing. The BOR has also begun using CAT in their certification program.

Many testing programs may not have the sample sizes of the above two groups, but still want to use IRT in their testing program. Sample sizes of 1,000 and tests of at least 50 items are generally recommended for the two- and three-parameter logistic IRT models, but samples of only 200 and 20 items are sufficient for the one parameter model (Barnes & Wise, 1991). However, it is generally agreed that the one-parameter model is not robust to violations of the assumption of zero lower asymptote, that is, guessing introduces significant error in the estimation of the item

ability estimates. Unfortunately, guessing is common in multiple-choice tests given by most licensing programs. Barnes and Wise (1991) examined the characteristics of the one-parameter model with a fixed non-zero lower asymptote. They compared the three-parameter model, and two forms of a modified one-parameter model. In MOD-1 the lower asymptote was fixed at the reciprocal of the number of response options (1/A). In MOD-2 the lower asymptote was fixed at 1/A - .05. Using simulated data, they varied the sample size (50, 100, and 200 candidates) and test length (25 and 50 items). The quality of each model was evaluated by examining the correlation between the true ability parameters and their estimates, the root mean squared errors (RMSEs) and bias of ability estimates, correlations between difficulty parameters and their estimates, RMSEs and bias of difficulty values, and RMSEs of recovered item characteristic curves. The results indicated that for all models the accuracy of item estimates improved with the longer test length. Further, the modified one-parameter models were observed to have lower RMSEs than the unmodified one-parameter model (and the three-parameter model), but the correlations between true parameters and ability estimates were comparable for both modified and unmodified one-parameter models. Although the results slightly favored MOD-2, the authors concluded that both modified models could be used effectively for multiple-choice tests with sample sizes of 200 and test lengths of 50 items, and both were an improvement over the one- and three-parameter models when only small sample sizes are available.

Because of IRT's advantages, I suspect that it will continue to play an ever increasing role in the larger licensure examination programs in the areas of test development and CAT. And for those testing programs with moderate to small sample sizes, modified one-parameter models appear to provide an avenue for experiencing the benefits of IRT.

COMPUTERIZED TEST ADMINISTRATION

During the 1980s licensing tests began to be administered with computer assistance. The first variant of computer-based testing (CBT) to be introduced involved the presentation of a paper-and-pencil test on a video screen. Technical support can be provided from either a LAN or minicomputer with dumb terminals. Candidates respond by either using the keyboard or touching the screen. An alternative form of presentation involves the use of a hand-held computer with a touch screen, thereby negating the need for a keyboard. Other options may exist, but all involve the presentation of a standard paper-and-pencil test on the computer, termed the "electronic page turner" by Friedman (1993). He identified several potential advantages to computerized testing, for both candidates and the provider of the tests. Probably the most significant advantage of this form of presentation to both groups is test security. No hard copy of the examination is provided to the candidate, and several forms of an examination can be made available simultaneously at one or more testing sites. Secondarily, instantaneous scoring and reporting of examination results are available if all scorable items have been used before. Pretest items can be included for analysis, but are not scored. Finally, test content can be more easily updated.

An alternative form of computerized testing is computer-adaptive testing (CAT). Under this model, each candidate can receive a unique form of the examination, tailored to his or her level of expertise. A typical scenario follows. An item of medium level difficulty is presented to the candidate. If the candidate answers it correctly, a slightly more difficult item is presented. If the candidate answers an item incorrectly, a slightly less difficult item is presented. The examination continues in this fashion, with items presented near the current ability estimate, until the specified content is covered, and a suitable estimate of the candidate's ability is determined. Because every candidate theoretically can be administered a unique test form of variable length, determining when to stop the examination presents a potential problem. The most common stopping rules include (a) the presentation of examinations of fixed length, or (b) the determination of a candidate's ability within a specified precision estimate, usually after a minimum number of items have been presented in all required content areas. Although at first CAT was applied to educational populations, at least one certification and one licensing examination program have begun to administer computer-adaptive examinations. However, before implementation, several issues had to be examined.

One of the first considerations is that of the size of the item bank. In an effort to provide some guidance in this area, Stahl and Lunz (1993) studied the amount of overlap in examinations using CAT for various sizes of item pools. Data were examined from five different certification examinations, with item banks ranging from 183 to 823 items. One of their results confirmed an intuitive conclusion, indicating that larger item banks tend to have a lower percentage of overlap among candidates, regardless of candidate ability. However, examinees close in ability tend to have a higher percentage of overlapping items. Considering both the amount of overlap and candidate ability, they concluded that a minimum desirable item bank size would be approximately 400–500 items, and that banks with 600–800 items are desirable.

In a national pilot study, Bergstrom and Lunz (1992b) examined the psychometric, psychological, and social attributes of CAT using a national sample of 645 medical technology students. Over 700 items were calibrated using the Rasch model (1-PL), and used as the item database for the CAT examination. They examined several issues relating to using CAT for certification examinations. Certification examinations are commonly built using spiral omnibus procedures, with easier items presented at the beginning, and more difficult items presented later in the examination. Therefore, one of their studies involved the starting difficulty (difficult, medium, or easy) of the first item presented to candidates. They found no difference in the starting difficulty of the first item, thus, no advantage appears to exist for starting the test with an easy item. They also observed that no significant differences existed in examinee performance for CATs with 50%, 60%, or 70% probability of correct response. This is of practical significance in that many item pools developed for occupational testing are targeted in the 70% range, and no major modification will be necessary for their use in CAT programs to challenge the more able examinee with items in the traditional 50% probability range of correct response.

Two final results included the observation that examinees who were allowed to manipulate their test (skip, review, and defer items) performed significantly better than those who had no control over their CAT, and those candidates who were administered the written test first did better on the CAT, suggesting a practice effect. The authors concluded that CAT is a feasible method of certification testing, and that it will likely become an accepted method of test administration.

A study by Legg and Buhr (1992) evaluated examinee attitudes toward CAT from another perspective. They analyzed data collected on college students on three adaptive tests: reading, mathematics, and writing. The data were examined to determine if examinees with different demographic characteristics (age, gender, ethnicity, ability, and experience with computers) displayed different response patterns to a questionnaire about testing conditions. It is encouraging that few differences were observed among the examinee groups that could not be addressed by expanding the pre-exam practice time.

In the standard method of CAT, examinees are not allowed to review previously answered items. The rationale is that if an examinee alters a response to an earlier item, an inaccurate estimate of his or her ability may result. However, for many licensure and certification examination programs, candidates consider this review to be one of their "basic rights." Thus, non-review of items may be a major political obstacle to the use of CAT for an occupational testing program. Lunz, Bergstrom, and Wright (1992) examined the effect of reviewing previously administered items on the estimation of students' abilities. The sample consisted of a geographically diverse group of 712 medical technology students. They were administered items from a database designed to be consistent with the test specifications of a national certification program in medical technology. Items were calibrated using the Rasch model (1-PL). Students were randomly assigned to a review group (n=220) or a non-review group (n=492). Their results indicated that the ability estimates for the students in the review group were correlated .98 before and after review. This conclusion is important because many candidate populations in this arena might feel uncomfortable without the opportunity to review and possibly change previously answered items.

Numerous studies have shown that computerized adaptive tests (CAT) can reduce test length without loss of precision in estimating a candidate's ability. Bergstrom and Lunz (1992a) examined the effect of test length on pass/fail decisions when using both CAT and paper-and-pencil examinations. The sample consisted of 645 medical technology students from 238 educational programs across the country, who were eligible for the next administration of a national certification examination. Each student took a CAT from a large bank of items, calibrated using the Rasch model (1-PL). Two versions of a written test, one short (109 items) and one long (189 items), were built from the same bank of items and were administered to the sample in a paper-and-pencil version, approximately 2 months after the CAT versions. Both written tests were analyzed using a Rasch calibration program. Their results indicated that while no significant differences existed among the CAT and paper-and-pencil tests, more pass/fail decisions could be made with 90% confidence for shorter CAT (mean length of 67 items) than with

longer (189 items) paper-and-pencil tests. The authors concluded that the implementation of CAT can reduce test length and improve confidence in the accuracy of pass/fail decisions.

A caution to some of these conclusions is provided by Vale (1993). He is in agreement that IRT can result in better balanced individual tests, a basic requirement for providing computerized testing on a daily basis. However, it has been his experience that the discrimination indices typically found in most licensing and certification tests are not sufficiently high to justify the use of CAT. Additionally, he suggests that CAT is more appropriate for wide range measurement, typically found in scholastic assessment, and not for the dichotomous pass/fail decisions required in a licensing environment. Fortunately, the current decade should provide much empirical data on the use of CAT in licensing and certification examinations.

EMPIRICAL ITEM BIAS REVIEW

Item bias, in particular differential item functioning (DIF), is another issue that has a solid foothold in testing practices. The Mantel-Haenzel (Holland & Thayer, 1988) and IRT procedures are two popular techniques for investigating item bias. Although studies for licensing tests appear to be unpublished, Skaggs and Lissitz (1992) conducted an investigation of the consistency of item bias using different procedures across two forms of an eighth grade math test. They found the Mantel-Haenzel and the IRT methods to be the most consistent, but the degree of reliability was modest. A major conclusion was that more consistency existed for larger sample sizes (n=2,000), as opposed to smaller samples (n=600). Additionally, their study provided supportive evidence that when bias has been found, it is modest and tends to favor the minority group.

Swaminathan and Rogers (1990) investigated differential item functioning (DIF) using logistic regression procedures and Mantel-Haenzel. Using simulated data, they found that the logistic regression procedure was more powerful than Mantel-Haenzel for the detection of nonuniform DIF (when an interaction exists between ability level and group membership), and equally as powerful for detecting uniform DIF (when no interaction exists between ability level and group membership). Their study also supported the use of larger samples for DIF studies. They found a 75% detection rate for sample sizes of 250, and 100% detection for a sample size of 500. Perhaps the dearth of published item bias studies for licensing examinations is due to the lack of sufficient sample sizes. Only a handful of licensing programs test candidates in sufficient numbers that may provide focal groups samples of several hundred candidates (for example, the National Council of State Boards of Nursing). Although authentic assessment is designed to increase the job-relatedness of an examination, increased content validity does not preclude the presence of bias in the assessment instrument. A study by Zwick, Donoghue, and Grima (1993) addressed the topics of the application of DIF procedures to performance tests. As part of their study they applied two Mantel-Haenzel procedures to the assessment of male-female DIF in constructed response reading and writing items, collected from 2,000 eleventh grade examinees as part of the 1990

NAEP (National Assessment of Educational Progress) program. They concluded that dichotomous DIF procedures were feasible for polychotomous (constructed-response) items, but cautioned that DIF procedures are only one component of examining the validity and fairness of performance assessment.

The major stumbling block for empirical item bias procedures to many licensing and certification testing programs is that of sample size. As the studies above indicate, large sample sizes are needed to provide consistent results with accurate detection for either IRT or Mantel-Haenzel procedures. However, some IRT procedures have been examined that may allow for smaller sample sizes for one of the target groups. For example, Linn and Harnisch (1981) suggested an IRT approximation that examined the difference between expected probability of correct response and observed proportion correct for the focal group. DIF analyses using the Mantel-Haenzel procedure may prove to be the most usable for many testing programs because of its more modest sample size requirements and its relative ease of use when compared to IRT procedures.

BIAS PANEL REVIEW

Frequently the large samples necessary to conduct DIF studies are not available. An alternative to empirical bias studies is the use of "sensitivity review" panels. Mehrens and Popham (1992) suggest that every high-stakes test (one that is used for high-stakes decisions such as employment) be evaluated for content relevance and potential bias by a sensitivity review panel. This type of panel can be used when the focal group is not sufficiently large for meaningful DIF analysis (50 or more individuals). They suggest that the bias review committee have representatives of the major protected groups who will be taking the test, and all participants be thoroughly trained in the process.

A procedure for accomplishing this review may include the establishment of a bias review committee, preferably separate from the standard examination committee. This will eliminate the possibility that the reviewers may have been too actively involved in writing, modifying, and editing items to give them a truly "non-partisan" review. The main responsibility of this committee is to review each examination item for possible bias with respect to gender and/or ethnic background. Each individual would receive thorough training on the review procedure, and respond individually to the following questions (adapted from W. J. Popham, personal communication, April 19, 1993) for each item using a rating sheet. The first three questions develop evidence in support of content validity, and the last two relate specifically to potential bias.

1. Is the content of this item necessary for successful performance as an entry level practitioner?
2. Is the task, knowledge, or skill appropriately measured by this item?
3. Of all knowledge or skills that entry level practitioners need, what percentage is represented by this test? (This question is answered after review of the complete test.)
4. Is this item biased against people due to gender, ethnic background, and/or socioeconomic status?

5. Might this item offend or unfairly penalize anyone due to gender, ethnic background, and socioeconomic status?

The rating sheets are summarized for each test item, and in those instances where less than 80% of the participants approve of an item, the item is revised before future use or deleted from the item bank (Mehrens & Popham, 1992).

The above item review procedures are recommended for every test used for licensure and certification. They should be used at initial review of the first test form to identify items that may not be appropriate for the desired purpose of the test, or have the potential to discriminate unfairly against protected classes. Later, if the sample sizes are sufficient for calculation of DIF statistics, additional items may be flagged as problematic. These items should not be automatically removed from future test use merely because of statistical evidence, but subjected to the same thorough review by a representative group of content experts. If this review fails to identify an obvious reason for the bias, Popham and Mehrens (1992) recommend that they remain in the item bank for future use.

CONCLUSIONS

Every aspect of licensure testing will continue to evolve with new directions or advances in educational and psychological measurement. Refinements to existing job analysis procedures will be made as different univariate and multivariate statistical techniques are employed to summarize the data and develop test specifications. The computer will play an ever increasing role in test construction and administration, allowing the refinement of existing item formats and the use of a variety of new item formats. It is hoped the desirable characteristics of the multiple-choice and constructed-response formats will be combined into a new format that retains the best psychometric characteristics of multiple-choice, but allows the benefits of authentic assessment to be realized in a cost-effective manner. Research will continue in the area of standard setting. Future studies will be conducted that will provide a rationale for techniques that adhere to the necessary technical requirements but are cognizant of the political realities of determining passing points for licensure examinations. Item response theory will strengthen its foothold and become the standard procedure for licensure test development and analysis for many programs. Computer-based testing, either in standard or adaptive format, will increase in popularity, eventually replacing paper-and-pencil presentations for the larger examination programs. Increasing numbers of programs will employ bias review panels prior to test administration to minimize undesirable discrimination for protected classes. Where technically feasible, empirical item bias procedures will be employed after the examination is given to ensure increased fairness to all examinees. These technological refinements and advances will help licensure testing become more precise such that both agencies and candidates will benefit.

REFERENCES

Albanese, M. A. (1993). Type K and other complex multiple-choice items: An analysis of research and item properties. *Educational Measurement: Issues and Practice*, *12*(1), 28-32.

Andrew, B. J., & Hecht, J. T. (1976). A preliminary investigation of two procedures for setting examination standards. *Educational and Psychological Measurement, 36*, 45-50.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508-600). Washington, DC: American Council on Education.

Barnes, L. L., & Wise, S. L. (1991). The utility of a modified one-parameter IRT model with small samples. *Applied Measurement in Education, 4*(2), 143-157.

Bergstrom, B. A., & Lunz, M. E. (1992a). Confidence in pass/fail decisions for computer adaptive and paper-and-pencil examinations. *Evaluation and the Health Professions, 15*(4), 453-464.

Bergstrom, B. A., & Lunz, M. E. (1992b, April). Computer adaptive testing: A national pilot study. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores* (part 5, pp. 397-479). Reading, MA: Addison-Wesley.

Bloom, B., Englehart, M., Furst, E., Hill, W., & Krathwohl, D. (Eds.) (1956). *Taxonomy of educational objectives, Handbook 1: The cognitive domain.* New York: David Makay Company, Inc.

Breyer, F. J. (1993). The Beuk compromise adjustment: Possible Rx for troubled cut-score study results? *CLEAR Exam Review, 4*(2), 23-27. Lexington, KY: CLEAR.

Bridgeman, B. (1992). A comparison of quantitative questions in open-ended and multiple-choice formats. *Journal of Educational Measurement, 29*(3), 253-271.

Beuk, C. H. (1984). A method for reaching a compromise between absolute and relative standards in examinations. *Journal of Educational Measurement, 21*, 147-152.

Case, S. M., & Downing, S. M. (1989). *Performance of various multiple-choice item types on medical specialty examinations: Types A, B, C, K, and X.* Philadelphia: National Board of Medical Examiners.

Costin, F. (1970). The optimal number of alternatives in multiple-choice achievement tests: Some empirical evidence for a mathematical proof. *Educational and Psychological Measurement, 30*, 353-358.

Dawson-Saunders, B., Nungester, R. J., & Downing, S. M. (1989). *A comparison of single best answer multiple-choice items (A-type) and complex multiple-choice (K-type).* Philadelphia: National Board of Medical Examiners.

Downing, S. M. (1992). True-false, alternate-choice, and multiple-choice items. *Educational Measurement: Issues and Practice, 11*(3), 27-30.

Dryden, R. E., & Frisbie, D. A. (1975, April). *Comparative reliabilities and validities of multiple-choice and complex multiple nursing education tests.* A paper presented at the annual meeting of the National Council on Measurement in Education, Washington, D.C.

Ebel, R. (1972). *Essentials of psychological measurement*, pp. 492-494. Englewood Cliffs, NJ: Prentice-Hall.

Friedman, C. (1993, September). On your mark, get set, go! Are you ready to put away your pencils and computerize your testing program? In C. Friedman

(Chair), *Computer-Based Testing*. Symposium conducted at the annual meeting of the Council for Enforcement, Regulation, and Licensure, San Diego, CA.

Geisinger, K. F. (1991). Training judges to generate standard-setting data. *Educational Measurement: Issues and Practice*, *10*(2), 17-22.

Haladyna, T. (1992a). Context-dependent item sets. *Educational Measurement: Issues and Practice*, *11*(1), 21-25.

Haladyna, T. (1992b). The effectiveness of several multiple-choice formats. *Applied Measurement in Education*, *5*, 73-88.

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn (Ed.), *Educational measurement*, (pp. 147-200). New York: Macmillan.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer-Nijoff.

Harasym, P. H., Price, P. G., Brandt, R., Violato, C., & Lorscheider, F. L. (1992). Evaluation of negation in stems of multiple-choice items. *Evaluation & the Health Professions*, *15*, 198-220.

Hixon, S. J. (1985). *An investigation of the psychometric properties of a clinical simulation examination for respiratory care practitioners*. (Unpublished doctoral dissertation, Ohio State University, Columbus)

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenzel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Earlbaum.

Hubbard, J. P. (1978). *Measuring medical education: The tests and experience of the national board of medical examiners* (2nd ed.) Philadelphia: Lea and Febinger.

Jaeger, R. M. (1991). Selection of judges for standard-setting. *Educational Measurement: Issues and Practice*, *10*(2), 3-6, 10, 14.

Kolstad, R. K., Bryant, B. B., & Kolstad, R. A. (1983). Complex multiple-choice items fail to measure achievement. *Journal of Research and Development in Education*, *17*, 7-11.

Landrum, R. E., Cashin, J. R., & Theis, K. S. (1993). More evidence in favor of three-option multiple-choice tests. *Educational and Psychological Measurement*, *53*, 771-778.

Legg, S. M., & Buhr, D. C. (1992). Computerized adaptive testing with different groups. *Educational Measurement: Issues and Practice*, 11(2), 23-27.

Linn, R. L., & Harnisch, D. L. (1981). Interactions between item content and group membership on achievement test items. *Journal of Educational Measurement*, *18*, 109-118.

Livingston, S. A., & Zieky, M. J. (1982). *Passing scores*. Princeton, NJ: Educational Testing Service.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Earlbaum Associates.

Lord, F. M., & Novick, M. R. (1968). *Theories of mental test scores*. Reading, MA: Addison-Wesley.

Lunz, M. E., Bergstrom, B. A., & Wright, B. D. (1992). The effect of review on student ability and test efficiency for computerized adaptive tests. *Applied Psychological Measurement*, *16*(1), 33-40.

Mehrens, W. A., & Popham, W. J. (1992). How to evaluate the legal defensibility of high-stake tests. *Applied Measurement in Education*, *5*(3), 265-283.

Mills, C. N. (1983). A comparison of three methods of establishing cut-off scores on criterion-referenced tests. *Journal of Educational Measurement*, *20*, 283-292.

Mills, C. N., Melican, G. J., & Ahluwalia, N. T. (1991). Defining minimal competence. *Educational Measurement: Issues and Practice*, *10*(2), 7-10.

Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement*, *14*, 3-19.

Nettles, S. S. (1987). *Psychometric characteristics of complex multiple-choice items*. (Unpublished doctoral dissertation, Rutgers University, New Brunswick.)

Norcini, J., & Shea, J. (1992). The reproducibilty of standards over groups and occasions. *Applied Measurement in Education*, *5*, 63-72.

Owen, S. V., & Froman, R. D. (1987). What's wrong with the three-option multiple-choice items? *Educational and Psychological Measurement*, *47*, 513-522.

Plake, B. S., & Kane, M. T. (1991). Comparison of methods for combining the minimum passing levels for individual items into a passing score for a test. *Journal of Educational Measurement*, *28*(3), 249-256.

Poggio, J. P., Glasnapp, D. R., & Eros, D. S. (1981, April). *An empirical investigation of the Angoff, Ebel, and Nedelsky standard setting methods*. Paper presented at the annual meeting of the American Educational Research Association, Los Angeles.

Popham, W. J. (1993). Educational testing in America: What's right, what's wrong? *Educational Measurement: Issues and Practice*, *12*(1), 11-14.

Rasch, G. (1960). *Probabalistic models for some intelligence and attainment scores*. Copenhagen, Denmark: Nielson & Lydiche.

Reid, J. B. (1991). Training judges to generate standard-setting data. *Educational Measurement: Issues and Practice*, *10*(2), 11-14.

Rosenfeld, M., & Thornton, R. F. (1976). *A case study in job analysis methodology*. Princeton, NJ: Educational Testing Service.

Schweibert, L. P., Davis, A. B., & Jacocks, M. A. (1992). Reproducibility of oral exam grades and correlations with other measures of performance on three required third-year clerkships. *Evaluation & the Health Professions*, *15*, 221-230.

Shaefer, L., Raymond, M. , & White, A. S. (1993). A comparison of two methods for structuring performance domains. *Applied Measurement in Education*, *5*, 321-335.

Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement*, *30*, 215-232.

Shea, J., Norcini, J., Baranowski, R. A., Langdon, L. O., & Popp, R. L. (1992). A comparison of video and print formats in the assessment of skill in interpreting cardiovascular motion studies. *Evaluation & the Health Professions*, *15*, 325-340.

Sistrunk, F., & Smith, P. L. (1982). Multimethodological job analysis for criminal justice organizations. In J. V. Ghorpade, *Job analysis: A handbook for the human resource director*, (pp. 130-134). Englewood Cliffs, NJ: Prentice Hall.

Skaggs, G., & Lissitz, R. W. (1992). The consistency of detecting item bias across different test administrations: Implications of another failure. *Journal of Educational Measurement*, *29*, 227-242.

Skakun, E. N., & Kling, S. (1980). Comparability of methods for setting standards. *Journal of Educational Measurement, 17*, 229-235.

Stahl, J. A., & Lunz, M. E. (1993, April). *Assessing the amount of overlap among computerized adaptive tests.* Paper presented at the annual meeting of the National Council of Measurement in Education, Atlanta, GA.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*, 361-370.

Vale, C. D. (1993, September). Should computer-adaptive testing be used in a licensing program? In C. Friedman (Chair), *Computer-Based Testing*. Syposium conducted at the annual meeting of the Council for Enforcement, Regulation, and Licensure, San Diego, CA.

Wainer, H. (Ed.). (1990). *Computer adaptive testing: A primer.* Hillsdale, NJ: Lawrence Earlbaum Associates.

Wainer, H., & Thissen, D. (1993). Combining multiple-choice and constructed-response test scores: Toward a Marxist theory of test construction. *Applied Measurement in Education, 6*(2), 103-118.

Wright, B. D. (1968). Sample-free test calibration and person measurement (pp.85-101). In *Proceedings of the 1967 ETS Invitational Conference on Testing Problems*. Princeton, NJ: Educational Testing Service.

Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of DIF for performance tasks. *Journal of Educational Measurement, 30*, 3, 233-251.